Proposal: Intelligent In-Network Attack Detection

Problem Statement and Motivation

The rapid growth of cyberattacks has exposed the limitations of traditional defense mechanisms. Conventional approaches such as firewalls and intrusion detection systems are often reactive, slow to adapt, and consume valuable server resources. As attackers adopt more sophisticated methods, it has become clear that security solutions must evolve to be adaptive, intelligent, and capable of analyzing traffic at line rate.

At the same time, network port speeds in modern data centers are increasing much faster than the compute capacity of servers. This widening gap creates a performance bottleneck: while servers can only process a limited number of cycles, incoming data volumes continue to surge. As a result, systems must handle higher traffic loads while still ensuring real-time security monitoring and minimal latency.

SmartNICs and DPUs offer a promising path forward by enabling in-network acceleration. By integrating security pipeline components directly into programmable network hardware, we can offload computationally intensive tasks from servers, enable real-time packet analysis, and reduce overall latency.

Related Work

Recent advancements in edge technologies have motivated significant exploration of in-network inference using SmartNICs and programmable switches. Traditional intrusion detection systems running on CPU cores (e.g., Snort, Suricata, Zeek) suffer from data transfer bottlenecks and latency overheads, limiting their effectiveness in high-speed scenarios. To overcome these challenges, recent efforts focus on deploying ML pipelines directly in the data plane. Tasdemir et al. [1] show that classical ML models can run efficiently on NVIDIA BlueField DPUs for intrusion detection, achieving both high accuracy and low latency. Similarly, Kapoor et al. [2] introduce ML-NIC, a framework that maps ML models onto Netronome SmartNICs, demonstrating significant improvements in inference latency and throughput over CPU-based approaches.

Other works extend this line of research toward anomaly detection and programmable switch pipelines. Wu et al. [3] propose ONLAD-IDS, a semi-supervised system on BlueField DPUs that adapts to evolving traffic while maintaining high detection performance. Xavier et al. [4] explore translating decision tree classifiers into P4 pipelines for programmable switches, showing that accurate in-network classification is achievable at line rate. Together, these studies establish the feasibility of in-network ML inference for security tasks and motivate further exploration of SmartNIC-accelerated anomaly detection using both P4 pipelines and CPU-core execution.

Initial Hypotheses

- **Hypothesis 1:** Attack detection accuracy can be improved by combining machine learning-based anomaly detection with signature-based methods on SmartNICs.
- **Hypothesis 2:** Offloading detection logic to SmartNICs will reduce host CPU utilization and latency compared to traditional software IDS solutions.

Goals:

1. Design an in-network attack detection pipeline that extracts key traffic features and applies trained ML models for real-time classification.

2. Build a prototype using P4/DPDK or DOCA libraries on a SmartNIC/DPU.
3. Evaluate performance in terms of accuracy, throughput, latency, and resource usage compared to baseline IDS.

Proposed Solution

We propose an intelligent in-network attack detection framework that combines feature extraction, machine learning, and signature-based methods. The system extracts traffic features (e.g., flow duration, packet size distribution, inter-arrival times) and analyzes them using:

- A machine learning model trained on labeled data to recognize anomalous or malicious traffic patterns.
- A knowledge base containing predefined attack signatures for rule-based detection.

A decision module then evaluates severity and confidence, triggering either real-time blocking or alerts for investigation. This hybrid approach ensures adaptability to new threats while maintaining reliability against established ones.

Evaluation

The performance of the proposed in-network attack detection system will be evaluated using a combination of machine learning and system-level metrics. From the ML perspective, we will measure accuracy to capture overall correctness, precision to quantify the number of flagged flows that are truly malicious, and recall to assess how effectively actual attacks are detected. To balance these two, the F1-score will be reported as a harmonic mean of precision and recall, ensuring a fair evaluation across both false positives and false negatives. On the systems side, we will examine inference time, which reflects the speed at which the model processes features and produces results, and latency, which measures the end-to-end delay introduced into the traffic pipeline. Together, these metrics will provide a holistic view of both the detection quality and the practical feasibility of deploying the system in real-time high-speed network environments.

References

[1] K. Tasdemir, R. Khan, F. Siddiqui, S. Sezer, F. Kurugollu and A. Bolat, "An Investigation of Machine Learning Algorithms for High-bandwidth SQL Injection Detection Utilising BlueField-3 DPU Technology," 2023 IEEE 36th International System-on-Chip Conference (SOCC), Santa Clara, CA, USA.

[2] Kapoor, R., Anastasiu, D. C., & Choi, S. (2025). ML-NIC: accelerating machine learning inference using smart network interface cards. *Frontiers in Computer Science*, *6*, 1493399.

[3] M. Wu, H. Matsutani and M. Kondo, "ONLAD-IDS: ONLAD-Based Intrusion Detection System Using SmartNIC," *2022 IEEE 24th Int Conf on High Performance Computing & Communications; 8th Int Conf on Data Science & Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*, Hainan, China.

[4] B. M. Xavier, R. S. Guimarães, G. Comarela and M. Martinello, "Programmable Switches for in-Networking Classification," *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, Vancouver, BC, Canada.